

COMPANION WORKSHEET

Chapter 17 — Health Baseline

Use this worksheet after Chapter 16 is Ready for Chapter 17. It records one health snapshot and one frozen-prompt baseline. It does not delay existing protective copies.

Work from a fresh copy. Use TBD, Evidence missing, or Not applicable.

A. Freeze the measured object

Chapter 16 worksheet path

Chapter 16 verdict

must be Ready for Chapter 17

Chapter 16 generation artifact identity actually used

That identity still belongs to the Chapter 14 operating set

primary / applicable fallback / other recorded role

Inherited runtime identity

Inspectable artifact identity

not latest

Timing path

must be direct inherited /api/generate

Knowledge / RAG / web search / tools attached to the timed prompt

must be No

Frozen prompt text, written before send

Generation settings actually sent

at least num_ctx or recorded defaults

Pre-run ollama ps names

Unrelated models stopped with ollama stop

Yes / inherited concurrent set preserved / Evidence missing

Why any remaining concurrent set is the inherited Chapter 5 workload, or Not applicable

If Chapter 16 is not Ready for Chapter 17, stop. If Chapter 16 used a fallback, time that identity. Do not silently time the primary.

B. Service, logs, and API

Current process and current API probe are the health evidence. The log is supporting context.

Current service or process state
present now / absent / Evidence missing

Official status command used

Official log source identified and inspected
Yes / Evidence missing

Recent log observation
no relevant recent line / normal startup line / error present

Current bounded API probe used

Current API result
reachable / error / Evidence missing

A chat box that answers is not this row. A historical log line is not current health. A missing health line is not, by itself, Evidence missing.

C. Resource snapshot

For each row: value, units, meaning, tool or method, timestamp, host label. Do not add unlike pools.

CPU observation

System memory observation

Accelerator or VRAM observation
or Not applicable

Disk free on the volume that holds models or the UI store

Temperature
value / Not applicable / unavailable by this lane’s official method

D. Frozen prompt timing

Send only the already-written prompt on the direct API. Use the Chapter 16 generation artifact.

Official fields — total_duration, load_duration, prompt_eval_count, prompt_eval_duration, eval_count, eval_duration — or Evidence missing

Units
official values are nanoseconds

Start state
unloaded-start / already-resident start, from pre-run ollama ps

Derived generation rate, if computed
 $\text{eval_count} / (\text{eval_duration} / 1\text{e}9)$ labeled as derived

Derived TTFT, if any — streaming wall-clock from request start to first generated chunk only / Evidence missing

Completion or stop reason

Why prompt_eval_duration is not TTFT

Why a fluent paragraph is not this measurement

E. Symptoms, not a diagnosis tree

Compare tonight to the Chapter 11 Processor record. Do not open Chapter 21.

Placement vs Chapter 11
same / changed / Evidence missing

Capacity symptom
none / memory pressure / disk pressure / Evidence missing

Memory-growth symptom
one row only / growth between comparable snapshots / Evidence missing

Fallback symptom
none / CPU fallback / error / Evidence missing

Changed placement is Blocked. Two snapshots can show growth. They do not prove a leak.

F. Assign one verdict

Allowed verdicts are exactly Ready for Chapter 18, Blocked, or Evidence missing.

Finite verdict

Binding reason

Next proof — Ready → Chapter 18; Blocked placement → Chapter 11 revalidation; otherwise named gap

Invalidation condition — include runtime, inspectable artifact identity, prompt, settings, loaded set, or placement change

Ready for Chapter 18 requires a named measured object that is the Chapter 16 generation artifact, a direct-API timing path with no Knowledge, frozen loaded-model state, current process and current API probe, a resource snapshot with named tools, matching Chapter 11 placement, and official usage fields. The log is supporting context. It is not a speed trophy.

Deterministic self-check

- ☐ Chapter 16 Ready for Chapter 17 was consumed, not repaired.
- ☐ The timed identity is the Chapter 16 generation artifact and still belongs to the Chapter 14 operating set.
- ☐ The timed path is direct /api/generate with no Knowledge, RAG, web search, or tools.
- ☐ Loaded-model state and generation settings were written before send.
- ☐ Current process and current API probe are the health evidence.
- ☐ The official log source was inspected; a missing health line is not automatically Evidence missing.
- ☐ Resource pools keep their labels and name their tools.
- ☐ Official usage fields were recorded; prompt_eval_duration is not called TTFT.
- ☐ Placement matches Chapter 11, or the verdict is Blocked / Evidence missing.
- ☐ Next proof matches the verdict.
- ☐ Exactly one allowed verdict and one invalidation condition are recorded.